A Three-pass System Combination Framework by Combining Multiple Hypothesis Alignment Methods

Jinhua Du
Centre for Next Generation Localisation
School of Computing, Dublin City University
Dublin, Ireland
jdu@computing.dcu.ie

Andy Way
Centre for Next Generation Localisation
School of Computing, Dublin City University
Dublin, Ireland
away@computing.dcu.ie

***Abstract*—So far, many effective hypothesis alignment metrics have been proposed and applied to the system combination, such as TER, HMM, ITER and IHMM. In addition, the Minimum Bayes-risk (MBR) decoding and the confusion network (CN) have become the state-of-the-art techniques in system combination. In this paper, we present a three-pass system combination strategy that can combine hypothesis alignment results derived from different alignment metrics to generate a better translation. Firstly the different alignment metrics are carried out to align the backbone and hypotheses, and the individual CN is built corresponding to each alignment results; then we construct a super network by merging the multiple metric-based CN and generate a consensus output. Finally a modified consensus network MBR (ConMBR) approach is employed to search a best translation. Our proposed strategy outperforms the best single CN as well as the best single system in our experiments on NIST Chinese-to-English test set.**

***Keywords*-three-pass; system combination; hypothesis alignment; super network;**

I. Introduction

In the past several years, multiple system combination has been shown to be helpful in improving translation quality. Recently, confusion network-based networks in [1], [2], [3], [4], [5], have become the state-of-the-art methodology to implement the combination strategy. A CN is essentially a directed acyclic graph which is built by a set of translation hypotheses against a reference or "backbone". Each arc between two nodes in the CN denotes a word or token, possibly a *null* item, with an associated posterior probability.

Typically, the dominant CN is constructed on the word level by a state-of-the-art framework. Firstly, a minimum Bayes-risk (MBR) decoder [6] is utilised to choose the backbone from a merged set of hypotheses, and then the remaining hypotheses are aligned against the backbone by a specific alignment approach. Currently, most of the research in system combination has focussed on the hypothesis alignment due to its significant influence on combination quality. TER-based [7] system combination strategy was firstly introduced in [3]. In recent years, many hypothesis alignment metrics have been proposed and successfully applied in system combination, such as ITG [8] and IHMM [5] etc. In all these papers, the proposed alignment method outperformed the TER-based baseline system.

A multiple CNs or super network framework was firstly proposed in [9], which used each of all individual system results as the backbone to build the CNs based on the same alignment metric – TER. A consensus network MBR (ConMBR) approach was presented in [3] which employs an MBR decoding to select the best one with the minimum cost from the original single system outputs compared to the consensus output. In this paper, we propose an idea that employs the MBR, super network and a modified ConMBR to construct a three-pass system combination framework which can effectively combine different hypothesis alignment results and easily be extended to more alignment metrics.

The remainder of this paper is organised as follows. In section II, we summarize three mainstream hypothesis alignment metrics–TER, HMM and IHMM, which have different working mechanism and represent the current state-of-the-art metrics in system combination. Section III introduces the modified ConMBR (mConMBR) decoding. Then, the Section IV describes the implementation details of our proposed three-pass combination strategy which combines the three different hypothesis alignment metrics. The experiments conducted on NIST Chinese-to-English data sets are reported in Sections V and VI. Section VII concludes and gives our future work.

II. Summary of Three Hypothesis Alignment Metrics

Hypothesis alignment is essentially an optimization problem on word alignment. The objective function is to search a best path of word alignment links between the source sentence F and the target sentence E.

A. TER

The TER (translation error rate) metric measures the ratio of the number of edit operations between the hypothesis E' and the reference E_b to the total number of words in the E_b. Here the backbone E_b is assumed as the reference. The allowable edits include insertions (Ins), deletions (Del), substitutions (Sub) and phrase shifts (Shft). The TER of E' compared to E_b is computed as

$$TER(E', E_b) = \frac{Ins + Del + Sub + Shft}{N_b} \times 100\%$$

where N_b is the total number of words in E_b. We can see that if the *Shft* is not permitted, the TER turns to Word Error Rate (WER) measure.

The TER is originally developed as an evaluation metric not an alignment metric. The $Shft$ edit is carried out by a greedy algorithm and restricted by three constraints: 1) The shifted words must match the reference words in the destination position exactly. 2) The word sequence of the hypothesis in the original position and the corresponding reference words must not exactly match. 3) The word sequence of the reference that corresponds to the destination position must be misaligned before the shift [7].

B. HMM

HMM-based hypothesis alignment model was presented in [2]. The idea is to consider alignment between the backbone sentence and the hypothesis sentence as a hidden variable in the conditional probability $P_r(E'|E_b)$. Given the backbone sentence $E_b = \{e_1, \ldots, e_I\}$ and the hypothesis sentence $E' = \{e'_1, \ldots, e'_J\}$, which are the same language, the alignment A between E_b and E' is defined as:

$$P_r(E'|E_b) = \sum_{A} P_r(E', A|E_b) \tag{1}$$

where $A \subseteq \{(j, i) : 1 \leq j \leq J; 1 \leq i \leq I\}$, i and j represent the word position in E_b and E' respectively. Hence, the alignment issue is to seek the optimum alignment $\hat{A}$ such that:

$$\hat{A} = \arg\max_{A} P(A|e_1^I, e_1'^J) \tag{2}$$

For HMM-based model, the Equation 1 can be represented as

$$P_r(e_1'^J|e_1^I) = \sum_{a_1^J} \prod_{1}^{J} [p(a_j|a_{j-1}, I) \cdot p(e'_j|e_{a_j})] \tag{3}$$

where $p(a_j|a_{j-1}, I)$ is the alignment probability and $p(e'_j|e_i)$ is the translation probability.

The model parameters are trained iteratively using the GIZA++ toolkit [10] which utilises the maximum likelihood estimation (MLE). The training is performed in the directions $E' \rightarrow E_b$ and $E_b \rightarrow E'$. The final alignment can be determined using the cost matrices [2] or the method of symmetrising–called refined method [10].

C. IHMM

IHMM-based (Indirect Hidden Markov) hypothesis alignment model was proposed in [5] which provides a different way to estimate the synonym matching and word ordering compared to the HMM method. In this approach, the parameters of the alignment model are estimated indirectly from a variety of functions, which use an interpolated similarity model p_{sim} to compute the translation probability $p(e'_j|e_i)$ and a distance-based distortion model p_d to obtain the alignment probability $p(a_j|a_{j-1})$. Therefore, the IHMM model can be written as

$$P_r(e_1'^J|e_1^I) = \sum_{a_1^J} \prod_{1}^{J} [p_d(a_j|a_{j-1}, I) \cdot p_{sim}(e'_j|e_{a_j})]$$

The similarity model is a linear interpolation model derived based on both semantic similarity p_{sem} and surface similarity p_{sur}:

$$p(e'_j|e_i) = \alpha \cdot p_{sem}(e'_j|e_i) + (1 - \alpha) \cdot p_{sur}(e'_j|e_i)$$

where the p_{sem} is calculated via the bi-directional lexical probabilities between the foreign words and the target words, and the p_{sur} is obtained using the longest matched prefix (LMP) algorithm to measure the string similarity. α is the smooth factor.

The distortion model estimates the first-order dependencies of word ordering, which assumes the alignment probabilities $p(a_j|a_{j-1})$ depend only on the jump distance $(i - i')$ [11]:

$$p(i|i') = \frac{c(d)}{\sum_{l=1}^{I} c(l - i')} = \frac{c(i - i')}{\sum_{l=1}^{I} c(l - i')} \tag{4}$$

where $\{d = i - i' : -4 \leq d \leq 6\}$ indicates the distortion parameter.

III. Modified Consensus Network MBR Decoding

In order to retain the coherent phrases in the original translations [3], it is sometimes better to retain sentence level consensus rather than creating new word-level consensus which may distort the fluency of the translation. This approach is defined as ConMBR. Firstly, the consensus network decoding is performed to obtain the combination result E_{con}. Then, the hypothesis in the original translations which has the minimum risk loss w.r.t. E_{con} is chosen as the consensus output, that is,

$$\hat{E}_{conMBR} = \arg\min_{E'} L(E', E_{con}) \cdot P(E|F) \tag{5}$$

where $L(E', E_{con})$ is the loss function under a specific evaluation metric. $P(E|F)$ is the posterior probability, usually set to a uniform distribution or can be trained as a system weight by a normalisation process.

However, it is believed that some of the new generated consensus sentences are better than the original ones. Inspired by the above way, we consider to merge these combination results from the different CNs with the original translations and then use the MBR decoder to research a best result. Hence, this method is defined as a modified form of ConMBR (mConMBR).

The NIST BLEU-4 [12] is used as the loss function in mConMBR which is computed as

$$\begin{aligned} L_{BLEU}(E', E) &= 1 - BLEU(E', E) \\ &= 1 - \exp\big(\frac{1}{4} \sum_{n=1}^{4} \log p_n(E', E)\big) \cdot \gamma(E', E) \end{aligned}$$

where $p_n(E', E)$ is the precision of n-grams in the hypothesis E' given the reference E. $\gamma \in [0, 1]$ is a brevity penalty.

Therefore, our mConMBR can be rewritten as

$$E_{mconMBR} = \arg\min_{E'} \big(1 - BLEU(E', E_{con})\big) \quad (6)$$

Here we set the posterior probability $P(E|F)$ to be a uniform distribution.

IV. Three-pass System Combination Strategy

A. Motivation

In recent years, many hypothesis alignment metrics have been proposed using different ways to solve the word alignment issue. The idea of a multiple CNs was presented in [9] which only uses TER as the alignment metric. Considering that the different hypothesis alignment links could bring different combination results, so we intend to use the combination techniques to combine multiple alignment metrics to improve the translation quality. There are two crucial contributions in our proposed method: 1) we are trying to use the diverse alignment results derived from different hypothesis alignment metrics in a unified combination framework; 2) we integrate the super network and mConMBR to combine these alignment metrics and fully make use of the translation results to improve the final quality.

B. Description of Algorithm

In sentence level, the different hypothesis alignment could produce different alignment results. See Figure 1 as an illustration. In Figure 1(a), E_b is the backbone selected from the MBR decoding, E_1 and E_2 are the original hypotheses from different machine translation (MT) systems. Fig. 1(b)(c)(d) show part of the alignment results performed by TER, HMM and IHMM respectively. We can find that the word "$america$" is misaligned to the word "$blood$" by TER in Fig. 1(b) while it is correctly aligned to "$american$" by HMM in Fig. 1(c) and by IHMM in Fig. 1(d). It is hard to automatically recognize and evaluate which alignment is better. In order to make full use of the different alignment results and increase the diversity of the searching process, we try to combine them in a super network. An example joint network with the priors for each metric and with votes for each arc are shown in Figure 2. According to the word alignment performed by a specific metric, an individual CN can be built with the voting or posterior probability on each arc as shown in Fig. 2.

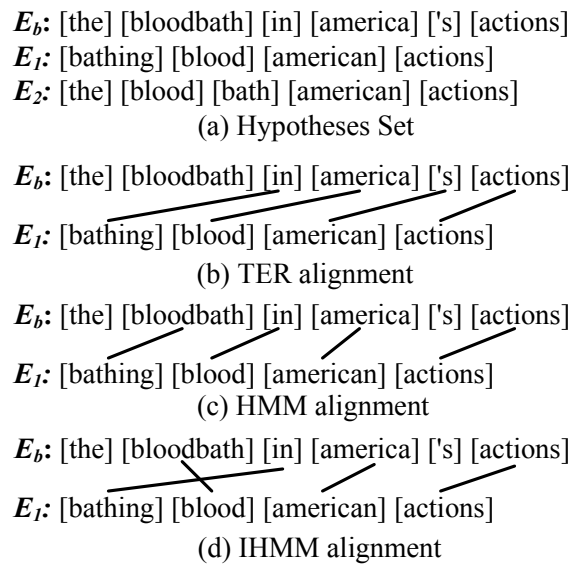

Figure 1. Hypotheses set and the word alignments performed by different metrics

In Fig. 2, the super network is constructed by integrating the TER-based, HMM-based and IHMM-based individual CN with prior probabilities. The prior probability is manually estimated in light of the performance of each single network. eps in Fig. 2 is ϵ that indicates the $null$ arc. In our experiments, the HMM outperforms the other two metrics and the TER is a slightly better than IHMM in terms of BLEU score, so the weights for the three single networks are set to 0.5, 0.3 and 0.2 respectively. All the three CNs are connected to a single start node S of ϵ arcs which contain the prior probabilities. Meanwhile, the three CNs are ended by a link of ϵ arc to a common end node E. The final arcs have a probability of one.

The construction of the three-pass combination framework may be summarized as follows:

Pass 1: Specific Metric-based Single Network:

1) Merge all the hypotheses from single MT systems into a new N-best list N_s;
2) Utilise the standard MBR decoder to select one from the N_s as the backbone;
3) Perform the word alignment between the backbone and the other hypotheses via the TER, HMM and IHMM metrics respectively;
4) Carry out the word reordering based on the word alignment (TER has finished the reordering in the process of scoring) and build three individual confusion networks named as CN_{ter}, CN_{hmm} and CN_{ihmm};
5) Decode the three single networks and export the consensus outputs separately.

Pass 2: Super Network:

1) Referring to the 5th step in Pass 1, we train CN_{ter}, CN_{hmm} and CN_{ihmm} through a development set (devset) to get the weights of each metric-based network, and then estimate the prior probability for each network;
2) Connect the three networks by a start node and an end node to form a multiple hypothesis alignment-based CNs;
3) Decode the super network and generate a consensus output.

Pass 3: mConMBR:

1) Combine the N_s with the results from CN_{ter}, CN_{hmm} and CN_{ihmm} and the result from super network to build a new N-best list N_{con};

2) Use mConMBR decoding to search a best final result from N_{con}.

V. Experimental Settings

In this section, we introduce the experimental settings for evaluating and comparing our three-pass alignment-based framework on Chinese-to-English (C2E) pair.

A. Chinese-English Test Data

We trained 5 MT systems to obtain a set of translations. All the MT systems are phrase-based engine, so in order to produce different results with less correlation, we have to train some diverse translation models.

Diversity has a significant influence on the performance of system combination [13]. In order to increase the diversity, we sample the training data to train different translation models. Furthermore, we can adjust the parameters such as the distortion limit or use different devsets to reduce any such correlation.

5 sub-training data sets are randomly sampled from a large-scale database, each of which contains 400K sentence pairs, including the HK, ISI parallel data, UN and other news data.

The devset used for translation system parameter training is NIST MT05 test set which contains 1082 sentences; the devset used for system combination parameter tuning (including MBR decoding tuning, CN tuning) is NIST MT06 test set which contains 1664 sentences. The test set is the NIST MT08 "current" test set which has 1357 sentences from two different domains, namely newswire and web-data genres. All the dev and test sets have 4 references per source sentence.

In this task, all the results are reported in BLEU, NIST and Meteor scores. The parameters and weights in combination process are also optimized under the BLEU score.

VI. Experimental Results and Analysis

A. Results

Table I first shows the performance of the best and the worst single systems as well as the Oracle result in terms of the BLEU score. In this task, the HMM-based method achieved the best performance in these three individual CNs. The consensus outputs from the **Super_CN** and the **mConMBR** (which is the final output of the three-pass framework) respectively demonstrate a significant improvement by 4.24%, 3.26%, 0.65% and 1.39%, 1.09%, 0.25% relative points in terms of the BLEU, NIST and MTR compared to the HMM-based single network. Moreover, the **mConMBR** also significantly outperforms the **Super_CN**.

Table I
Results on Chinese-to-English Test Set

System	BLEU	NIST	MTR
Worst Single	17.33	6.59	39.82
Best Single	21.64	6.94	42.95
Oracle	26.67	7.93	44.95
TER-based	22.47	7.36	43.11
IHMM-based	22.45	7.34	43.20
HMM-based	**23.10**	**7.37**	**43.27**
Super_CN	**23.42**	**7.45**	**43.38**
mConMBR	**24.08**	**7.61**	**43.55**

B. Analysis

From the comparison results, we can find that the multiple-pass combination strategy achieved a significant improvement compared to the individual CN and the best single system.

Different hypothesis alignment metrics can bring different alignment results, which will increase the diversity of the searching process. Although this might increase the error risk of misalignment, we can see from the experiments, due to the close performance between the three individual CNs, it would not cause serious risk. On the other hand, it can provide more potentially correct candidates for the decoder to determine a final path. Such an intrinsic way that combines different hypothesis alignment results constructs a multiple word lattice networks, which can fully make use of the context information. Regarding the **mConMBR**, since the CN is built based on word-level, some new sentences could be generated and bring some new syntactic structures into the MBR decoding. Hence, the three-pass strategy based on the super network and the mConMBR are proved to be effective in our experiments.

VII. Conclusions and Future Work

In this paper, we investigated three dominant hypothesis alignment metrics used in system combination. Based on these metrics, we presented a unified three-pass framework to combine and utilise the alignment results so as to obtain an improved performance. We firstly run the word alignment between the backbone and the hypothesis using the TER, HMM and IHMM respectively and build the individual CN according to their respective alignment links, then connect these three networks with a common start node and a end node to form a super network. Finally, a modified ConMBR is carried out to search a best final translation from the N_{con} list. Experiments are conducted on Chinese-English langauge pair and the experimental results demonstrate the effectiveness of our proposed method.

As for future work, firstly we plan to automatically evaluate the alignment quality of different hypothesis alignment metrics. Secondly, we plan to examine how the differences between the hypothesis alignment metrics impact on the accuracy of the super network. We also intend to integrate more alignment metrics to the networks and verify on the other language pairs.

Acknowledgment

This work is supported by Science Foundation Ireland (Grant No. 07/CE/I1142). Thanks also to the reviewers for their insightful comments.

References

[1] S. Bangalore, G. Bordel and G. Riccardi, "Computing consensus translation from multiple machine translation systems," Proc. ASRU '01, 2001, pp. 351–354.

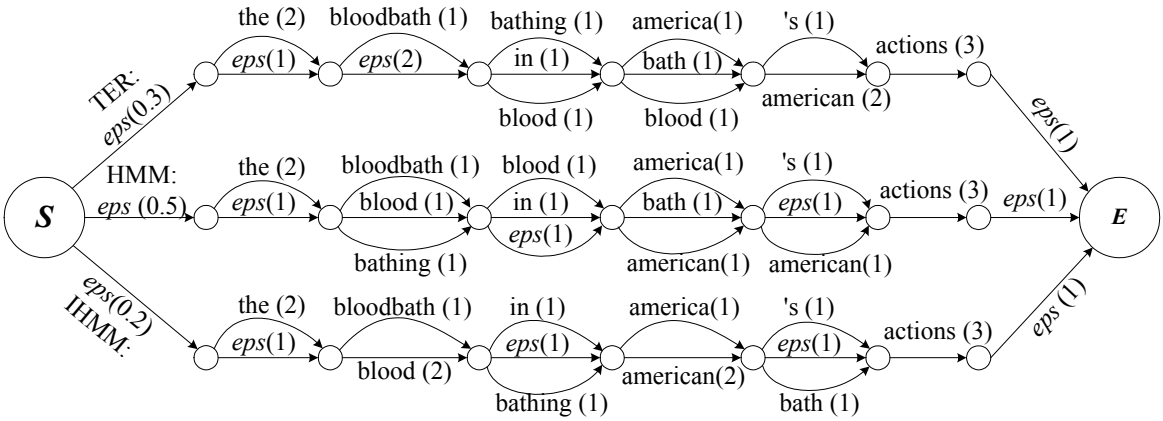

Figure 2. Hypothesis alignment-based multiple confusion networks with prior and posterior probabilities

[2] E. Matusov, N. Ueffing and H. Ney, "Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment," Proc. EACL'06, 2006,pp. 33–40.

[3] K.C. Sim, W.J. Byrne, M.J.F. Gales, H. Sahbi, and P.C. Woodland, "Consensus network decoding for statistical machine translation system combination," Proc. ICASSP, 2007, pp. 105–108.

[4] A.I. Rosti, B. Xiang, S. Matsoukas, R. Schwartz, N.F. Ayan and B.J. Dorr, "Combining outputs from multiple machine translation systems," Proc. HLT-NAACL, 2007, pp. 228–235.

[5] X. He, M. Yang, J. Gao, P. Nguyen and R. Moore, "Indirect HMM-based hypothesis alignment for combining outputs from machine translation systems," Proc. EMNLP'08, 2008, pp. 98–107.

[6] S. Kumar and W. Byrne, "Minimum Bayes-Risk Decoding for Statistical Machine Translation," Proc. HLT-NAACL, 2004, pp. 169–176.

[7] M. Snover, B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul, "A study of translation edit rate with targeted human annotation," Proc. AMTA, 2006, pp. 223–231.

[8] D. Karakos, J. Eisner, S. Khudanpur and M. Dreyer, "Machine translation system combination using ITG-based alignments," Proc. ACL'08, 2008, pp. 81–84.

[9] A.I. Rosti, S. Matsoukas and R. Schwartz, "Improved Word-Level System Combination for Machine Translation," Proc. ACL'07, 2007, pp. 312–319.

[10] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," Computational Linguistics, vol. 29, pp. 19–51.

[11] S. Vogel, H. Ney and C. Tillmann, "HMM-based word alignment in statistical translation," Proc. of the 16th International Conference on Computational Linguistics, 1996, pp. 836–841.

[12] K. Papineni, S. Roukos, T. Ward and W.J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," Proc. ACL'02, 2002, pp. 311–318.

[13] W. Macherey and F. Och, "An empirical study on computing consensus translations from multiple machine translation systems," Proc. EMNLP, 2007, pp. 986–995.